Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 32–36
Marseille, 20 June 2022

The Viability of Best-worst Scaling and Categorical Data Label Annotation Tasks in Detecting Implicit Bias

Parker Glenn[1,2], Cassandra L. Jacobs[1,3], Marvin Thielk[1], and Yi Chu[1]
1: Workhuman Inc., 2: Brandeis University, 3: University at Buffalo

Abstract

Annotating workplace bias in text is a noisy and subjective task. In encoding the inherently continuous nature of bias, aggregated binary classifications do not suffice. Best-worst scaling (BWS) (Louviere and Woodworth, 1991) offers a framework to obtain real-valued scores through a series of comparative evaluations, but it may be impractical to deploy to traditional annotation pipelines within industry. We present analyses of a small-scale bias dataset, jointly annotated with categorical annotations and BWS annotations and show that there is a strong correlation between observed agreement and BWS score (Spearman's r=0.72). We identify several shortcomings of BWS relative to traditional categorical annotation: (1) When compared to categorical annotation, we estimate BWS takes approximately 4.5x longer to complete; (2) BWS does not scale well to large annotation tasks with sparse target phenomena; (3) The high correlation between BWS and the traditional task shows that the benefits of BWS can be recovered from a simple categorically annotated, non-aggregated dataset.

Keywords: categorical annotation, best-worst scaling, scalability

1. Introduction

Social bias, or the preference for one class of people over another, is pervasive in our day-to-day interactions with the world. *Implicit bias* in language occurs when producers intentionally or unintentionally reveal their beliefs about a person or a group of people. The field of natural language processing has taken significant strides to eliminate biases from text (Bolukbasi et al., 2016; Garg et al., 2018; Zhao et al., 2017), though it is clear that producers are the source of the ultimate biases observable within corpora (Trix and Psenka, 2003; Blair, 2002).

In the present work, we present a novel data source from workplace interactions between employees in the same organization. The dataset contains instances of "social recognition" in which an employee (e.g., the *author*) praises a coworker (the *recipient*), such as for obtaining a career milestone such as a promotion, or for completing a difficult task successfully. We focus specifically on *workplace bias*, which we define as any language that detracts from the general positivity of the praise, such as instances of discrimination (e.g., "You're a great engineer for a woman!"), the promotion of unhealthy work-life balance (e.g., "Thanks for working until nighttime."), or self-centered praise.

However, because the bias is implicit, the linguistic phenomena that reflect workplace bias show considerable degrees of subjectivity. Along the *intersubjectivity spectrum* (Basile et al., 2021), annotating for specific categories of workplace bias relies considerably on the annotators' existing conceptualization of the classes. This makes the categorical labels effectively "projective latent content", lacking clear boundaries even with strong guidelines (Reidsma and op den Akker, 2008). In the context of bias identification, approaching the annotation process with a sense of distrust or hyperfocus on the raters' abilities to "correctly" annotate can quickly lead to erasure of diverse and valuable opinions in how bias is received (Basile et al., 2021). Indeed, it is partly due to this distrust that analyses of annotated data often aggregate across many raters, reducing the contribution of any single individual's biases.

In an effort to encode some of the nuance associated with highly subjective social phenomena, researchers have used Best-Worst Scaling (BWS) (Louviere and Woodworth, 1991) as one popular approach. (Mohammad, 2017; Pei and Jurgens, 2020). Kiritchenko and Mohammad (2017) verified the efficacy of BWS by obtaining judgments of positivity and negativity for 3,207 terms using both a 9-point rating scale and the BWS framework. They showed that using BWS produced more reliable annotations than rating scales (Kiritchenko and Mohammad, 2017).

We explore the potential viability of the BWS annotation procedure, which has been proposed in contrast to categorical data labeling in domains such as word affect intensity (Mohammad, 2017), intimacy (Pei and Jurgens, 2020), hate speech (Poletto et al., 2019), and sentiment (Kiritchenko and Mohammad, 2017), which are similar to our workplace bias dataset. However, the nuance of workplace bias makes it a distinct annotation problem, posing its own unique set of difficulties when implementing BWS at scale.

2. Methodology

As we aim to evaluate the viability of the BWS annotation procedure compared to traditional categorical labeling, we compile a jointly annotated dataset in both styles and analyze the results.

2.1. Dataset

For our study, we compile 50 social recognition messages between co-workers at various companies. Social recognition, or peer-to-peer recognition, is the act of employees empowering and acknowledging one another for great work. The messages are shared on the Workhuman online platform, where employees from a company write these peer-to-peer messages. For example, the following recognition is found in our dataset: *I want to appreciate you for working together and collaborating as a team in difficult times*.

Four trained linguists annotators rated 50 social recognition messages at the sentence level, where each message had on average 4.5 sentences ($sd = 1.3$). In total, the 50 messages yielded 227 categorically annotated sentences, with an average of 18.4 tokens ($sd = 12$). Of these, 107 sentences received a positive bias annotation, indicating the presence of some workplace bias category by at least one annotator, and thus went on to be annotated in the style of Best-Worst Scaling (BWS). Only these 107 jointly annotated sentences are included in the dataset. To our knowledge, this is the first analysis directly comparing BWS to parallel non-aggregated categorical labels. Before we introduce the BWS annotation procedure, we discuss the categorical annotation procedure.

2.2. Categorical Annotation

In our taxonomy, we define six categories for classifying instances of workplace bias. It is similar in nature to the typology of microaggressions described by Breitfeller et al. (2019). However, our implicit bias annotation task centers on nuanced language specific to the workplace. For legal reasons, we anonymize the bias categories in the present study to be of the form "Category {id}", in addition to "None" (the absence of any gold standard workplace bias category).

For each of the 227 sentences, annotators may identify multiple categories applying to a single sentence. The resulting average Fleiss' κ statistic across all categories is 0.32. This value represents low agreement in categorical annotation, but must be framed in the context of other difficult implicit bias annotation tasks, reporting κ values as low as 0.43 (Breitfeller et al., 2019)

For each datapoint, we find a single "gold standard" category by aggregating judgements and taking the most common category, where more than one annotator (at least 50% of the annotators) selected that category.

2.3. Best-Worst Annotation

Best-worst Scaling (BWS) is a method of annotation in which a series of comparative judgements are aggregated in order to produce real-valued scores corresponding to some criteria (Louviere and Woodworth, 1991). Rather than performing binary comparisons between all pairs of items in a dataset (N^2 complexity), the items are grouped into "tuples" of four datapoints, leveraging the transitivity property to maximize the information gained for each evaluation item. In our context, the criterion in question is "bias potency": how strong is the bias present in a given text? Being a subjective task capturing projective latent content, this annotation paradigm is intentionally ambiguous and category-agnostic, resulting in a lower cognitive overhead. However, our working definition of potency is a measure of the negative impact a text will have on both workplace culture and the individual recipient.

The final BWS scores are obtained using Counts Analysis (Orme, 2009). For each item a, the score is calculated as follows:

$$bws\_score(a) = \%best(a) - \%worst(a) \qquad (1)$$

The final bws_score ranges from -1 (least potent workplace bias) to 1 (most potent workplace bias). An example of a BWS annotation item is shown in Figure 1.

We cannot apply traditional inter-annotator agreement algorithms like Alpha and Kappa to the set of BWS annotations, since all forms of disagreement will be penalized. However, disagreement that comes from two items having similar ratings is a useful signal in BWS, since these two items will ultimately be pushed towards having more similar real-valued scores (Mohammad, 2017). Instead, we calculate the split-half reliability correlation to ensure that the levels of disagreement are replicable across many random splits of annotations. Across 100 random splits, these tests yield a Spearman's r of 0.84, demonstrating high reliability in the annotations.

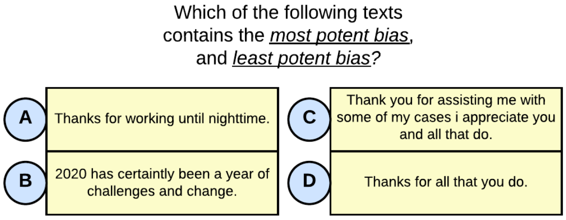
Which of the following texts
contains the most potent bias,
and least potent bias?
A
Thanks for working until nighttime.
B
2020 has certaintly been a year of challenges and change.
C
Thank you for assisting me with some of my cases i appreciate you and all that do.
D
Thanks for all that you do.

Figure 1: Example BWS item.

3. Analysis

Observed Agreement as Substitute for BWS

When calculating inter-annotator agreement on a traditional, categorically annotated dataset, a simple non-chance corrected metric used is observed agreement. Observed agreement is traditionally defined simply by the proportion of cases in which two raters agree. In our context, we slightly adapt this definition to reflect the direction of agreement, such that observed

Category	N	Coefficient	σ
None	47	-0.55	0.29
Category 1	11	0.25	0.36
Category 2	19	0.33	0.32
Category 3	11	**0.12**	**0.40**
Category 4	13	**0.03**	**0.36**
Category 5	1	0.13	0.0
Category 6	4	0.06	0.38

Table 1: Point biserial coefficients between category and BWS score, alongside standard deviations of BWS scores. In bold are notable examples where $N > 10$, and $\sigma - |\text{coefficient}| > 0.3$

agreement is *(number of annotators who identified any bias category in an item) / (total number of annotators for an item)*. The resulting values fall within the range of [0, 1], where 1 implies that all annotators agreed that some form of bias is present, and 0 implies that all annotators agreed that no bias is present.

Spearman's r between the observed agreement and the BWS scores is 0.72, demonstrating a strong positive correlation. Figure 2 plots a regression model fit between observed agreement scores and BWS scores.

Predicting Bias Potency

In examining the taxonomy of bias annotated for in the categorical annotations, it might seem fair to assume that certain categories inherently carry more bias potency than others. However, although the bias categories are indeed annotated as having higher degrees of bias than the "None" (no gold standard) category, there are no notable differences in the bias potency of different categories, which we show in Figure 3. Table 1 calculates the point biserial correlations and standard deviations with respect to each aggregated category, and "None". From the inferred confidence intervals in this table, it is clear that estimating bias potency through category alone is insufficient.

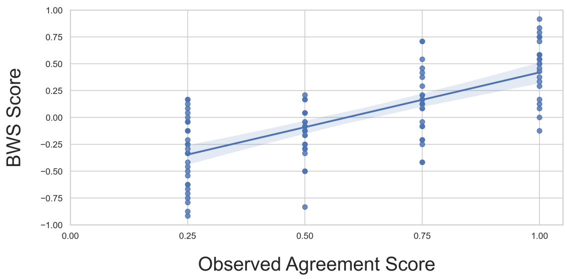

Figure 2: Regression plot between observed agreement and mean BWS scores. Note: Agreement scores of 0 were not used in the BWS annotation.

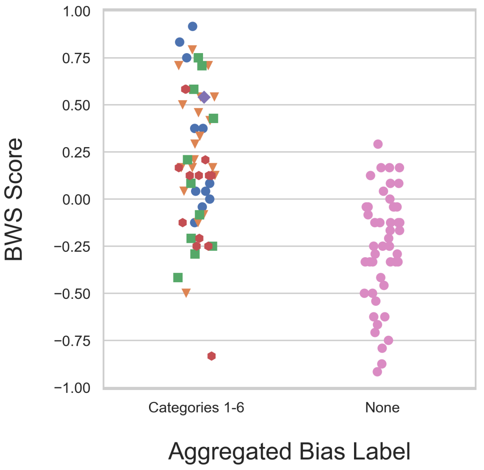

Figure 3: "Gold Standard" bias categories, plotted against BWS score. The hues correspond to specific bias categories.

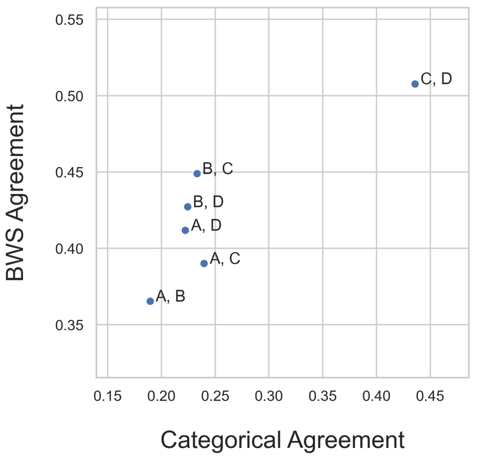

Figure 4: Agreement percentages for the BWS and categorical tasks between pairs of annotators A, B, C, and D.

In order to examine the extent to which mental conceptions of implicit bias impact agreement on the BWS and categorical task, we plot the observed agreement between pairs of annotators in Figure 4. For categorical annotations, agreement is the proportion of cases where the two raters' categorical judgements align. For BWS scaling, this is the proportion of best/worst judgements in which two annotators agree.

Annotating Sparse Phenomena with BWS

As seen in Figure 2, there are no observed agreement values equal to 0. This is a result of the data preprocessing we performed prior to BWS annotation; any sentence receiving less than two categorical annotations indicating the presence of some form of bias was discarded. This pre-processing is motivated by both practical annotation constraints and confounding linguistic considerations.

An internal dataset of 4,224 sentences from Workhuman social recognition data shows that only 452 (12%) of the sentences shows instances of perceived bias, according to our taxonomy. When confronted with a tuple of all-unbiased sentences, the random disagreement amongst annotators will likely produce relatively similar scores. However, annotating in the style of BWS when 88% of the data contain none of the target phenomena is costly and creates a massive overhead.

Surveying other applications of BWS in annotating social aspects of language shows that others do not employ similar pre-processing prior to BWS annotation. For example, Pei and Jurgens (2020) use BWS to annotate Reddit questions on intimacy levels, defined as the perceived independence, warmth, and willingness to share personally (Perlman and Fehr, 1987). For a social phenomenon as ubiquitous as intimacy, this lack of preprocessing might work well. However, for the annotation of a sparse phenomenon like bias, many datapoints will completely lack the trait in question (bias). Indiscriminate annotation of all datapoints in the BWS style may lead to unwanted priming, resulting in more positive annotations due solely to the rating environment or noisy linguistic cues in the prompt (Schuster et al., 2019). As shown in Figure 5, we observe a discrepancy between the percentage of data points that received the lowest score even with our preprocessing filtration, with 11.32% in the BWS task, and 33.02% in the categorical task. As a result, we see that BWS slightly skews judgements towards the biased side, similar to the conclusions made in Poletto et al. (2019).

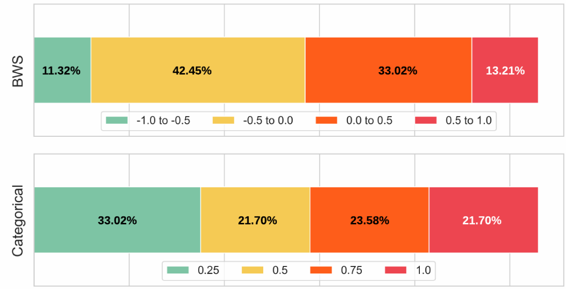

Figure 5: Label distribution for BWS and categorical annotation tasks.

Scaling to New Data

To annotate in the BWS style, a complete and final set of the data is required. If dataset A is annotated for bias potency in a BWS style, it may be realistic that dataset B becomes available at a later date. Since Figure 3 shows the difficulty in predicting distribution bias potency, there is no clear way to ensure the datasets are drawn from the same distribution of bias potency. As a result, a new annotation job must be created, based on a composite dataset of A + B.

This example highlights a major downside to deploying BWS at scale in an industry setting: the final scores are only interpretable in context of a reference dataset. However, the observed agreement of each data point is interpretable in isolation and shows a high correlation with the BWS scores.

Annotation Times

Additionally, annotation time must be considered when choosing a large-scale annotation pipeline. In the categorical annotation task, the annotators spent an average of 25 seconds per sentence. Annotating in the BWS style, annotators report spending an average of 60 seconds per tuple. These discrepancies in annotation times are highlighted when we consider the relative sizes of the datasets for annotation. In order to construct well-formed tuples which produce meaningful scores, the number of tuples is commonly made to be (at minimum) 1.5 times the size of the original dataset, though we note that Kiritchenko and Mohammad (2017) show that the reliability of BWS annotations is similar across 1N, 1.5N, and 2N tuples. As a result, the average annotator spent ~ 45 minutes annotating 107 sentences in a categorical paradigm and ~ 3.5 hours annotating 107 sentences in a BWS paradigm. While the overhead of training annotators to annotate in the categorical style must be considered, this substantial difference in annotation times makes categorical annotations better suited for scaling an annotation pipeline.

4. Conclusion and Future Work

In this work, we analyzed the relationship between non-aggregated categorical annotations and BWS annotations. Analyses of a novel dataset of workplace bias showed a strong correlation between observed agreement and the BWS score. Given the often-unmentioned pitfalls of annotation time and complications annotating on sparse social phenomena, we propose leveraging categorical annotations as a more realistic alternative for perspectivist modeling approaches. Additionally, we demonstrate the value of non-aggregated datasets.

We hope to see more datasets jointly annotated in this manner so that our results might be validated on a larger scale. In future work, we hope to leverage the observed agreement scores from non-aggregated categorical bias annotations to inform a form of soft loss learning Basile et al. (2021).

5. Acknowledgements

We would like to thank Kristen Frazier, Kaan Catalyurek, Mimi Zhang, Cassie Newland, and Cassidy Copeland, all with Workhuman, for their annotations on the dataset and invaluable conceptualization and development of the workplace bias categories.

6. Bibliographical References

Basile, V., Cabitza, F., Campagner, A., and Fell, M. (2021). Toward a perspectivist turn in ground truthing for predictive computing. *ArXiv*, abs/2109.04270.

Blair, I. V. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, 6(3):242–261, August. Publisher: SAGE Publications Inc.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Breitfeller, L., Ahn, E., Jurgens, D., and Tsvetkov, Y. (2019). Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, April. Publisher: Proceedings of the National Academy of Sciences.

Kiritchenko, S. and Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

Louviere, J. J. and Woodworth, G. G. (1991). Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Mohammad, S. M. (2017). Word Affect Intensities. *arXiv:1704.08798 [cs]*, April. arXiv: 1704.08798.

Orme, B. (2009). Maxdiff analysis: Simple counting, individual-level logit, and hb. *Sawtooth Software*.

Pei, J. and Jurgens, D. (2020). Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020*.

Perlman, D. and Fehr, B. (1987). The development of intimate relationships.

Poletto, F., Basile, V., Bosco, C., Patti, V., and Stranisci, M. (2019). Annotating hate speech: Three schemes at comparison. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–8. CEUR-WS.

Reidsma, D. and op den Akker, R. (2008). Exploiting 'Subjective' Annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK, August. Coling 2008 Organizing Committee.

Schuster, S., Chen, Y., and Degen, J. (2019). Harnessing the linguistic signal to predict scalar inferences. *arXiv preprint arXiv:1910.14254*.

Trix, F. and Psenka, C. (2003). Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society*, 14(2):191–220, March.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September. Association for Computational Linguistics.